Temporary Staffing Services: A Data Mining Perspective

Submitted for Blind Review (DO NOT ADD AUTHORS' NAME IN BLIND REVIEW SUBMISSION)

***Abstract*— Research on the temporary staffing industry discusses different topics ranging from workplace safety to the internationalization of temporary labor. However, there is a lack of data mining studies concerning this topic. This paper meets this void and uses a financial dataset as input for the estimated models. Bagged decision trees were utilized to cope with the high dimensionality. Two bagged decision trees were estimated: one using the whole dataset and one using the top 12 predictors. Both had the same predictive performance. This means we can highly reduce the computational complexity, without losing accuracy.**

Data mining; Temporary staffing services; Bagged decision trees; Feature selection

I. INTRODUCTION

In employment literature the externalization of labor is often contrasted with the internalization of it. Both have a different purpose, but they work in a complementary way [1]. Internalization is focused on stability, while externalization is employed to improve flexibility [1;2]. This internalization-externalization dualism is also conceptualized as a "make-and-buy" strategy [3]. Here, the make strategy corresponds to internalization where a company builds a skilled employee base itself through the means of training and development, whereas in a buy strategy this process is outsourced [4]. Companies often use a combination of both strategies as each has its own specific costs and benefits [4;5].

Externalization is a term that refers to a broad range of non-standard working arrangements such as temporary workers, leased workers and independent contractors [1].The focus of this paper lays within the field of temporary employment. Temporary employees can be defined as those "Individuals who work at the establishment but who are paid through an employment agency and are not on the organization's payroll" [6, p. 151]. Nowadays, the temporary staffing services are considered big business [7]. Its corresponding revenue is relatively high, despite a generally low market penetration [8]. It is estimated that in 2005 the temporary staffing industry was already worth over €157 billion per year, and it continued growing.

In the past, temporary staffing was forbidden by law in some countries or discouraged by international conventions in others [8;9]. This changed, starting in the 1970s, gathering momentum in the 1980s and really expediting in the 1990s [8]. Belgian data clearly shows this relatively strong rise during the 1990s (Fig. 1). More specifically, in 1990 5.3% of the working force was employed in the temporary staffing industry, rising to 10.2% in 1999 [10]. Starting in 2000 there was a small decline in temporary workers. In 2010, Belgium had 8.1% temporary workers compared to a European percentage of 13.9.

A change is arising in the well-established markets, as these markets are becoming increasingly concentrated [7].

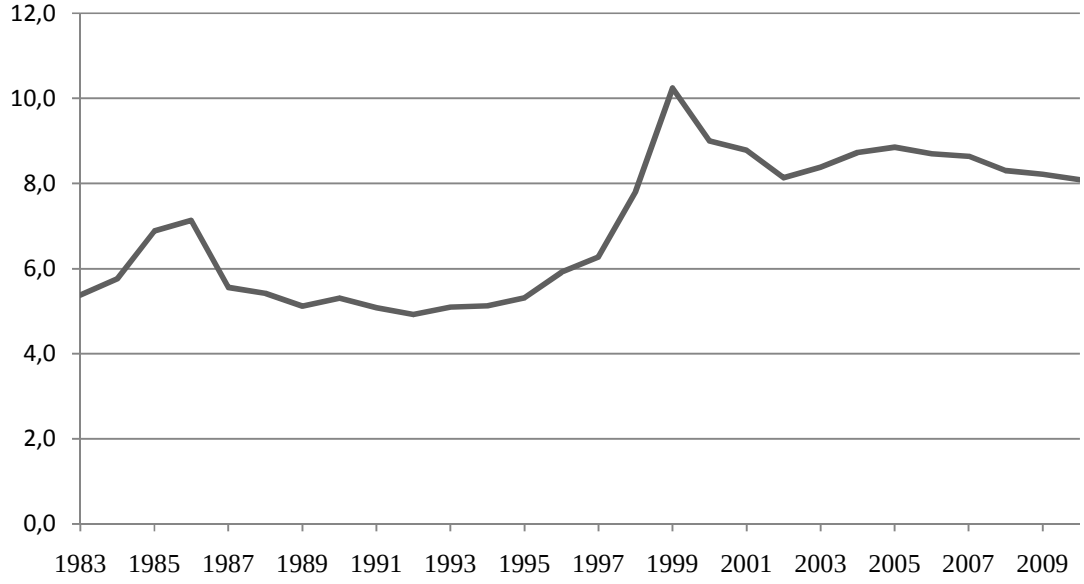

Figure 1. Percentage of temporary workers (15-64 year) in Belgium [source: 10]

A small group of big temporary staffing agencies are currently dominating the market. The biggest future market growth is expected in countries that were strongly regulated in the past, but are sturdily liberalizing now: e.g. Italy and Japan [8]. Temporary staffing supplied by agencies was only legalized in 1994 in Italy, giving rise to more than 300 agencies in a time span of three years. Liberalization is going much slower in Japan, but the potential growth in the long run is even greater. In Belgium the temporary employment market is more or less stabilized.

Research on the temporary staffing industry discusses different topics ranging from workplace safety [e.g. 11] to the internationalization of temporary labor [e.g. 12]. However, to our knowledge, there has not been conducted any data mining studies in this field. Data mining is a way of extracting knowledge hidden in large databases (Ngai et al, 2009). As the size of databases keeps growing, this type of analysis is becoming more and more important (Ngai et al, 2009; Rygielski et al, 2002). This paper tries to fill the data mining void and utilizes a bagged decision tree to make the predictions. Furthermore, a feature selection is done to greatly reduce the size of the bagged decision trees, while retaining its predictive performance. If the results show that the estimated models render a high accuracy in predicting the use of temporary staffing, they are of a high value to companies that offer temporary employees to other companies. These temporary staffing companies can use the models to predict whether new companies might be inclined to make use of their services. Especially in the Belgian context where there is a saturated market these types of models are valuable.

The remainder of the paper is structured as follows. First, the literature on temporary workers is reviewed. Next, we discuss bagged decision trees and the evaluation criterion. Then, the data is presented and the results are discussed. Finally, we end with a conclusion and discussion.

II. TEMPORARY WORKERS

Different explanations have been provided of why companies use temporary workers instead of regular employees. A first explanation is that some employers prefer temporary workers to lower employment costs such as training and monitoring costs or even wages [1;2;13;14]. For example, in some cases the wages for employees are set by a union contract and higher than the market average [6;13]. A company could as a result decide to use temporary workers that are not subject to this contract and pay them less. A second reason for hiring temporary workers is numerical flexibility [2;13;14]. It gives companies the ability to adjust to the variable demand in working force, for example in seasonality bound companies such as fruit growers. Building a regular staff large enough to meet the demand at peak moments is an inefficient way of working [13]. These additional employees would be non-active during slow moments, but they still have to be paid. Thirdly, temporary employees can be used to screen for potential regular workers [2;6;15;16]. It is often much easier to lay off a temporary worker compared to a regular worker. Legal issues prevent companies from simply firing regular employees. A selection of temporary workers can be hired and best candidate can be consequently hired as a regular employee, ending the contract of the other temporary employees. Unionization is seen as a possible cause as well, but the direction of the relationship is not clear [2]. It might prevent or stimulate the use of temporary staffing. The number of employees also has an influence on the use of temporary employees because larger companies can have the need for specialized services for which it is not efficient to produce them in-house [2;14]. A final reason for the use of temporary workers in a company is the ratio male-females [2]. A higher proportion of females is linked to more temporary employment. The reason is family obligations for women that makes it hard for them to combine full time working with a family.

Data mining techniques, nomen est omen, start from data to discover knowledge. Thus, we will not test specific hypotheses, as outlined above. A different angle is used. Instead of building explanatory models, we focus on building predictive models.

III. BAGGED DECISION TREES

Logistic regression is an often used and well-known data mining technique. However, it is not optimal in fitting models that have high dimensional datasets as input. Instead it is recommended using different techniques such as decision trees. A problem with a decision tree is that it has been shown to be unstable [17]. This means that small changes in the training data (e.g. a different random selection) can cause large changes in the predictions. A method to overcome this instability is bagging, short for bootstrap aggregating, developed by Breiman [18]. Bagging can be formalized as follows (Breiman, 1996a; Cunningham et al, 2000):

$$\hat{y}_{BAG} = \frac{1}{B}\sum_{b=1}^{B} \emptyset(x; T_b) \qquad (1)$$

where B is the number of bootstrap samples of training set T and x is the input. $\hat{y}_{BAG}$ is the average of the different estimated trees Fildes [19]. A bootstrap sample is randomly drawn from the training set, but with replacement [18]. Therefore, each observation can appear more than once in a single bootstrap sample or even not at all. The size of a bootstrap sample is usually chosen to be the same size as the training set [20]. It is important that when building bagged trees, the different trees are not pruned [21]. This is necessary because variability is needed for the averaging to give a stable result. A bootstrap sample leaves out about 37% of the

observations in the training data [18]. There is no general rule as to how many bootstrap samples should be used. Breiman [18] found that in his case, 50 were enough, while 100 did not decrease the accuracy. That is why we decided to take 100 bootstrap samples. As each bootstrap sample is random, a bagged tree will be different each time it is estimated. An additional advantage of bagged decision trees is that they are capable of outputting a measure of importance for the different variables. This is done by randomly permuting the values of the different variables and evaluating what the effect is on the predictions. Variables that have a high impact on the predictions after the random permutation are deemed more important than those who have a lower impact.

IV. Evaluation criterion

Performance measures are a crucial part of the analysis. A model with a low predictive power has no use, so a measure of model quality is necessary.

The confusion matrix represents the relation between the predicted and real values (Table I). A company that is predicted as using temporary staffing services and used it in real life as well is called a True Positive (TP). One that is predicted as being a non-user and did not use it in reality is called a True Negative (TN). A company that is predicted as using the services, but is not using them is called a False Positive (FP). A company predicted as not using temporary staffing services, but that is in fact using it is called a False Negative (FN). The following equations are a measure of quality:

$$Accuracy = PCC = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

		Predicted Status	
		User	Non-user
True Status	User	**True** Positive (TP)	**False** Negative (FN)
	Non-user	**False** Positive (FP)	**True** Negative (TN)

Table I. Confusion Matrix

PCC (2) stands for the percentage correctly classified and it calculates exactly that. Sensitivity (3) is equal to the true positive rate. Specificity (4) is equal to the true negative rate. The goal of your model is to get TP and TN as high as possible and FP and FN as low as possible. To be able to construct the confusion matrix, a cut-off value needs to be specified. This cut-off value defines which observations are predicted as using the services (= 1) and which are predicted as not using them (= 0), because most models
output a probability (i.e. a continuous number between 0 and 1, with 1 being the highest probability). The cut-off value can be a probability you define above which you categorize an observation as a user or it could be a real top proportion of your data that is considered as a user. However, this is also the main weakness of these evaluation criteria. There is no clear-cut way in deciding which is the ideal cut-off. As a result a more general performance measure is preferred.

The area under the receiver operating curve (also know as the 'AUC') is calculated to evaluate the overall quality of a model. AUC is a common metric to estimate the accuracy of a model [22]. It can vary from 0.5 to 1, with 0.5 being a random model and 1 being the perfect model [23]. The advantage of the AUC measure is that it is cut-off independent. It measures the performance over all possible cut-off values. It represents the probability that a randomly chosen positive example (a user) is ranked higher than a randomly selected negative example (a non-user). The AUC is in fact the relation between the sensitivity and 1-specificity. In the case of AUC being 0.5 your are equally likely to produce false positives as true positives. To conclude whether two AUC outcomes are statistically different we use a method that was developed by DeLong et al [24]. This method uses a Chi-Square to assess if two AUC results are significantly different.

V. Data

A database was used that contains financial data on Belgian companies. The database contains, among others, a summary of financial strength ratios, the average number of employees, key company financials, … A selection of 452 relevant variables were made. The main selection criterion was the amount of missing values. If the amount of missing values is too high, these variables are not included. Missing values are a big and often occurring quality problem in commercial databases. If missing values were present in the selected variables they were imputed using a tree imputation with surrogates. For each variable that had to be imputed an imputation dummy indicator was created resulting in a final dataset of 554 independent variables. This dummy has the value of one when an observation is imputed for a certain variable and the value of zero when it is not imputed. The reason for this is that missing values might have

predictive power as well. In some cases a missing value has a substantive interpretation. The dependent variable indicated whether a company made use of temporary staffing services or not. Around 70% of the company did not use temporary staffing and 30% did use it. The goal is to build a model that predicts the use of companies of temporary staffing services. All analyses were done using Matlab and SAS. Matlab was used to estimate the different bagged decision trees, while SAS was used to do the data preparation and calculating and comparing the performance of the different decision trees (AUC).

VI. Results

The first bagged decision tree rendered an AUC of 0.7586. This includes the full dataset of 554 variables. As mentioned above, an advantage of bagged decision trees is that they are able to output a measure of importance for each variable. The top variables were selected that had an importance higher than 0.5. This value was arbitrarily chosen, a different cut-off value would also be possible. As a result, 12 variables were retained that had an importance higher than the chosen value. A new bagged decision tree was estimated using these selected variables and it rendered an AUC of 0.7487, which is only marginally lower than the first bagged decision tree. Furthermore, this bagged decision tree was not statistically significant from the previous one, χ^2 (1, N = 4369) = 2.3749, p = 0.1233. This means that it can be concluded that both bagged decision trees have the same performance, but the second one is greatly reduced in size. This leads for example to a reduction in computing time. Estimating a bagged tree with the full set of variables takes about 1 hour, while estimating a bagged tree that only contained the top 12 predictors ran within minutes.

The top explanatory variables can be roughly divided into three categories. The first category comprises variables related to the financial health of the company (e.g. the yearly turnover of the company). The second category covers general variables concerning the employee base (e.g. the number of employees). The final

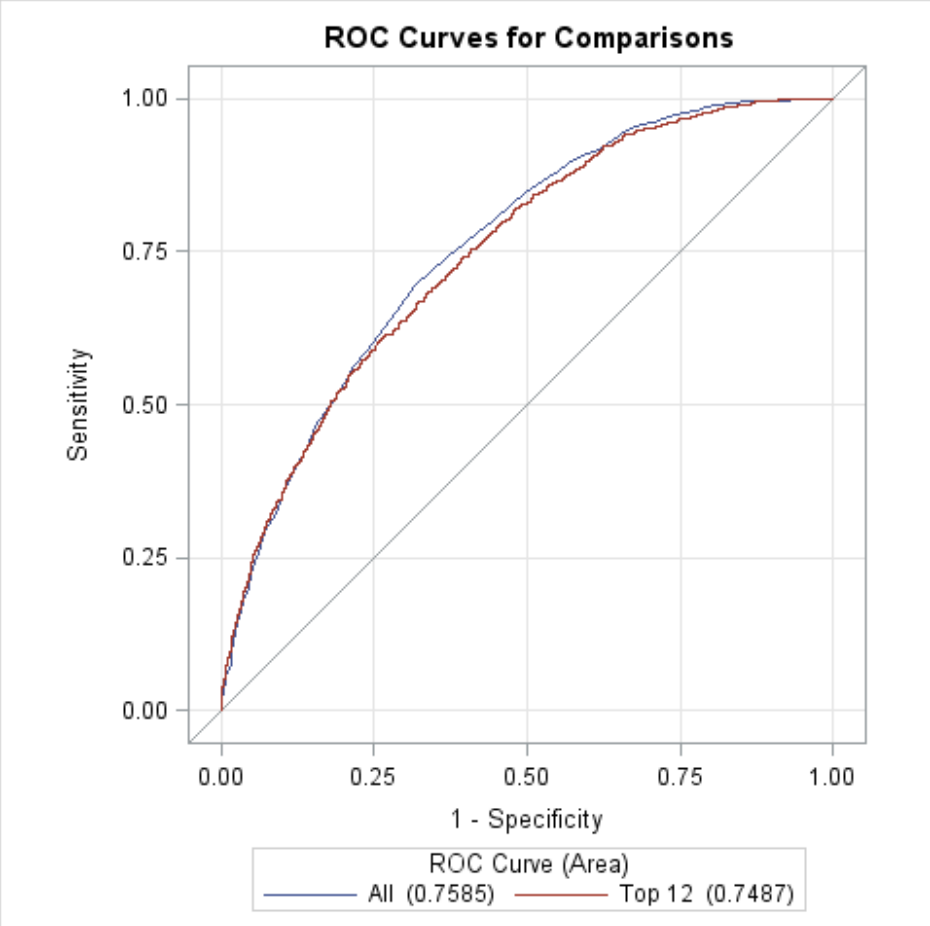

Figure 2. ROC Curves for Comparisons

category encompasses financial variables concerning the employee base (e.g. staff costs).

VII. Conclusion and discussion

There is a vast literature on temporary staffing services. Research on the temporary staffing industry discusses different topics ranging from workplace safety [e.g. 11] to the internationalization of temporary labor [e.g. 12]. However, as to date, no studies have been done using a data mining methodology in this field of study. The temporary staffing market is more or less stabilized in Belgium. This makes it hard for companies that offer temporary staffing services to other companies to grow in this saturated market. In this regard there is a need for studies that use a data mining point of view. Building models that are able to predict which new companies a temporary staffing provider should pursue is crucial for such a saturated market.

This paper tries to fill the data mining void uses a financial dataset as input for the estimated models. Logistic regression is an often used and well-known data mining technique. However, it is not optimal in fitting models that have high dimensional datasets as input. As a result, bagged decision trees were utilized to cope with this high dimensionality. Bagging decision trees was preferred instead of normal decision trees as the latter are shown to be unstable. Two bagged decision trees were estimated: one using the whole dataset and one using the top 12 predictors. The method created by DeLong et al [24] was used to check if both models performed significantly different. However, it was concluded that both bagged decision trees did not differ significantly in their performance. This means we can highly reduce the computational complexity, without losing accuracy. This results, inter alia, in a reduction of computational time. Estimating a bagged tree with the full set of variables takes about 1 hour, while estimating a bagged tree that only contained the top 12 predictors ran within minutes.
Furthermore, it is possible to conclude which are the best predictors of using temporary staffing. This way, temporary staffing providers know which variables they need to focus on and they do not need to use a high dimensional dataset that contains a high volume of irrelevant variables.

The top explanatory variables can be roughly divided into three categories. The first category comprises variables related to the financial health of the company (e.g. the yearly turnover of the company). The second category covers general variables concerning the employee base (e.g. the number of employees). The final category encompasses financial variables concerning the employee base (e.g. staff costs). This makes it clear that there are company specific variables that are linked to the use of temporary staffing services.

References

[1] A. Davis-Blake and B. Uzzi, "Determinants of Employment Externalization: A Study of Temporary Workers and Independent Contractors," *Administrative Science Quarterly*, vol. 38, no. 2, pp. 195-223, 1993.

[2] A. L. Kalleberg, J. Reynolds, and P. V. Marsden, "Externalizing employment: flexible staffing arrangements in US organizations," *Social Science Research*, vol. 32, no. 4, pp. 525-552, Dec.2003.

[3] R. E. Miles and C. C. Snow, "Designing strategic human resources systems," *Organizational Dynamics*, vol. 13, no. 1, pp. 36-52, 1984.

[4] D. P. Lepak and S. A. Snell, "The Human Resource Architecture: Toward a Theory of Human Capital Allocation and Development," *The Academy of Management Review*, vol. 24, no. 1, pp. 31-48, Jan.1999.

[5] C. Jiang and D. Cheng, "The Alignment of Internalization and Externalization Employment Mode on Performanec: The Case of Manufacturing Firms in the People's Republic of China," 2004.

[6] S. N. Houseman, "Why employers use flexible staffing arrangements: Evidence from an establishment survey," *Industrial and Labor Relations Review*, vol. 55, no. 1, pp. 149-170, 2001.

[7] K. Ward, "Going global? Internationalization and diversification in the temporary staffing industry," *Journal of Economic Geography*, vol. 4, no. 3, pp. 251-273, June2004.

[8] J. Peck, N. Theodore, and K. Ward, "Constructing markets for temporary labour: employment liberalization and the internationalization of the staffing industry," *Global Networks*, vol. 5, no. 1, pp. 3-26, 2005.

[9] A. L. Kalleberg, "Nonstandard Employment Relations: Part-Time, Temporary and Contract Work," *Annual Review of Sociology*, vol. 26, pp. 341-365, Jan.2000.

[10] FOD Economie, "Aandeel tijdelijke arbeid bij de loontrekkenden (15-64 jaar) naar geslacht in de Europese Unie, 1983-2010,"2011.

[11] C. Mehta and N. Theodore, "Workplace Safety in Atlanta's Construction Industry: Institutional Failure in Temporary Staffing Arrangements," *WorkingUSA*, vol. 9, no. 1, pp. 59-77, 2006.

[12] L. McDowell, A. Batnitzky, and S. Dyer, "Internationalization and the Spaces of Temporary Labour: The Global Assembly of a Local Workforce," *British Journal of Industrial Relations*, vol. 46, no. 4, pp. 750-770, 2008.

[13] K. G. Abraham and S. K. Taylor, "Firms' Use of Outside Contractors: Theory and Evidence," *Journal of Labor Economics*, vol. 14, no. 3, pp. 394-424, 1996.

[14] J. S. Heywood, W. Siebert, and X. Wei, "Estimating the Use of Agency Workers: Can Family-Friendly Practices Reduce Their Use?," *Industrial Relations: A Journal of Economy and Society*, vol. 50, no. 3, pp. 535-564, 2011.

[15] S. N. Houseman, A. L. Kalleberg, and G. A. Erickcek, "The Role of Temporary Agency Employment in Tight Labor Markets," *Industrial and Labor Relations Review*, vol. 57, no. 1, pp. 105-127, Oct.2003.

[16] A. Engellandt and R. T. Riphahn, "Temporary contracts and employee effort," *Labour Economics*, vol. 12, no. 3, pp. 281-299, June2005.

[17] L. Breiman, "Heuristics of Instability and Stabilization in Model Selection," *The Annals of Statistics*, vol. 24, no. 6, pp. 2350-2383, Dec.1996.
[18] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, Aug.1996.
[19] R. Fildes, K. Nikolopoulos, S. F. Crone, and A. A. Syntetos, "Forecasting and operational research: a review," *Journal of the Operational Research Society*, vol. 59, no. 9, pp. 1150-1172, May2008.
[20] G. Martinez-Munoz and A. Suarez, "Out-of-bag estimation of the optimal sample size in bagging," *Pattern Recognition*, vol. 43, no. 1, pp. 143-152, Jan.2010.
[21] R. A. Berk, "Bagging," in *Statistical Learning from a Regression Perspective* Springer Verlag, 2008, pp. 169-192.
[22] W. C. Chen, C. C. Hsu, and J. N. Hsu, "Optimal Selection of Potential Customer Range through the Union Sequential Pattern by Using a Response Model," *Expert systems with applications*, vol. 38, no. 6, pp. 7451-7461, June2011.
[23] P. Baecke and D. Van den Poel, "Data augmentation by predicting spending pleasure using commercially available external data," *Journal of Intelligent Information Systems*, vol. 36, no. 3, pp. 367-383, June2011.
[24] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 44, 837-845.